

Delta Theorem in the Age of High Dimensions*

Mehmet Caner

Department of Economics

Ohio State University

January 24, 2017

Abstract

We provide a new version of delta theorem, that takes into account of high dimensional parameter estimation. We show that depending on the structure of the function, the limits of functions of estimators have faster or slower rate of convergence than the limits of estimators. We illustrate this via two examples. First, we use it for testing in high dimensions, and second in estimating large portfolio risk. Our theorem works in the case of larger number of parameters, p , than the sample size, n : $p > n$.

*Department of Economics, 452 Arps Hall, TDA Columbus, OH, 43210. email:caner.12@osu.edu

1 Introduction

Delta Method is one of the most widely used theorems in econometrics and statistics. It is a very simple and useful idea. It can provide limits for complicated functions of estimators as long as function is differentiable. Basically, the idea is the limit of the function of estimators can be obtained from the limit of the estimators, and with exactly the same rate of convergence. In the case of finite dimensional parameter estimation, due to derivative at the parameter value being finite, rates of convergence of both estimators and function of estimators are the same.

In the case of high dimensional parameter estimation, we show that this is not the case, and the rates of convergence may change. We show that the structure of the function is the key, and depending on that functions of estimators may converge faster or slower than estimators. In this paper, we provide a new version of delta method which accounts for high dimensions, and generalizes the previous finite dimensional case. After that we illustrate our point in two examples: first by examining a linear function of estimators that is heavily used in econometrics, and second by analyzing the risk of a large portfolio of assets in finance. Section 2 provides new delta method. Section 3 has two examples. Appendix shows the proof.

2 High Dimensional Delta Theorem

Let $\beta_0 = (\beta_{10}, \dots, \beta_{p0})'$ be a $p \times 1$ parameter vector with an estimator $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$. Define a function $f(\cdot)$, $f : K \subset R^p \rightarrow R^m$ defined at least on a neighborhood of β_0 , where $p > m$, m will be taken as constant for convenience, but p is allowed to increase when n increases. Furthermore the function $f(\cdot)$ is differentiable at β_0 , which means, with $h \neq 0$,

$$\lim_{h \rightarrow 0} \frac{\|f(\beta_0 + h) - f(\beta_0) - f_d(\beta_0)h\|_2}{\|h\|_2} = o(1),$$

where $\|\cdot\|_2$ is the Euclidean norm for a generic vector, and $f_d(\beta_0)$ is the $m \times p$ matrix, which ij th cell consists of $\partial f_i / \partial \beta_j$ evaluated at β_0 , for $i = 1, \dots, m$, $j = 1, \dots, p$.

Before the theorem, we need the following matrix norm inequality. Take a generic matrix which is of dimension $m \times p$. Denote the Frobenius norm for a matrix as $\|A\|_2 = \sqrt{\sum_{i=1}^m \sum_{j=1}^p a_{ij}^2}$. Note

that in some of the literature such as Horn and Johnson (2013), this definition is not considered a matrix norm, due to lack of submultiplicativity. However, our results will not change regardless of matrix norm definitions, if we abide by Horn and Johnson (2013), our results can be summarized in algebraic form, rather than matrix norm format. Define

$$A = \begin{bmatrix} a'_1 \\ \vdots \\ a'_m \end{bmatrix},$$

where a_i is $p \times 1$ vector, and its transpose is a'_i , $i = 1, \dots, m$. Then for a generic $p \times 1$ vector x ,

$$\|Ax\|_2 = \sqrt{\sum_{i=1}^m (a'_i x)^2} \leq \left(\sqrt{\sum_{i=1}^m \|a_i\|_2^2} \right) \|x\|_2 = \|A\|_2 \|x\|_2, \quad (2.1)$$

where the inequality is obtained by Cauchy-Schwarz inequality. Note that if we apply Horn and Johnson (2013) norm definition, this matrix norm inequality still holds, but we cannot use the matrix norm. In that case we have

$$\|Ax\|_2 \leq \left(\sqrt{\sum_{i=1}^m \|a_i\|_2^2} \right) \|x\|_2. \quad (2.2)$$

Our new delta theorem is provided for high dimensional case. This generalizes Theorem 3.1 of van der Vaart (2000). Key element in our Theorem below is $\|f_d(\beta_0)\|_2$. We should note that this norm of matrix derivative depends on n , through p , which is the number of columns in $f_d(\beta_0)$. Let $r_n, r_n^* \rightarrow \infty$, as $n \rightarrow \infty$, be the rate of convergence of estimators, and the functions of estimators, respectively in the theorem below.

Theorem 2.1. *Let a function $f(\beta) : K \subset R^p \rightarrow R^m$, and differentiable at β_0 . Let $\hat{\beta}$ be the estimators for β_0 , and $\hat{\beta} \neq \beta_0$, assume we have the following result:*

$$r_n \|\hat{\beta} - \beta_0\|_2 = O_p(1).$$

a) Then if $\|f_d(\beta_0)\|_2 \neq o(1)$, with $\|f_d(\beta_0)\|_2 > 0$, we get

$$r_n^* \|f(\hat{\beta}) - f(\beta_0)\|_2 = O_p(1), \quad (2.3)$$

where

$$r_n^* = O\left(\frac{r_n}{\|f_d(\beta_0)\|_2}\right). \quad (2.4)$$

b) Then if $\|f_d(\beta_0)\|_2 = o(1)$, we get

$$r_n \|f(\hat{\beta}) - f(\beta_0)\|_2 = o_p(1). \quad (2.5)$$

Remarks. 1. Note that in part a), with r_n^* , we have a slower or the same rate of convergence as in r_n . In part b), clearly, the function of estimators converge to zero in probability faster than the rate of estimators themselves. This can be seen from noting that in part b), even though

$$\|\hat{\beta} - \beta_0\|_2 = O_p(1/r_n),$$

for functions

$$\|f(\hat{\beta}) - f(\beta_0)\|_2 = o_p(1/r_n).$$

2. Note that Horn and Johnson (2013) defines Frobenius norm only for square matrices unlike our case. If we use their approach, then our main result in part a) will be

$$r_n^* = O\left(\frac{r_n}{\sqrt{\sum_{i=1}^m \|f_{di}(\beta_0)\|_2^2}}\right), \quad (2.6)$$

where we use (2.2) instead of (2.1) in the proof of Theorem 2.1a. Also note that $f_{di}(\beta_0)$ is the $p \times 1$ vector, which is $\partial f_i(\cdot)/\partial \beta$ evaluated at β_0 .

3. Also see that this result (2.4) can be obtained in other matrix norms subject to the same

caveat in Remark 1. A simple Holder's inequality provides

$$\|Ax\|_1 \leq \|A\|_1 \|x\|_1, \quad (2.7)$$

where we define the maximum column sum matrix norm: $\|A\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^m |a_{ij}|$, where a_{ij} is the ij th element of A matrix. Applying this in the proof of Theorem 2.1a, given $r_n \|\hat{\beta} - \beta_0\|_1 = O_p(1)$ and replacing everything with Frobenius norm for matrices with maximum column sum matrix norm, and l_2 norm for vectors with l_1 norm, we have

$$r_n^* = O\left(\frac{r_n}{\|f_d(\beta_0)\|_1}\right). \quad (2.8)$$

Also part b) can be written in l_1 norm as well.

4. We can also extend these results to another norm. A simple inequality provides

$$\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty, \quad (2.9)$$

where we define the maximum row sum matrix norm: $\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^p |a_{ij}|$, where a_{ij} is the ij th element of A matrix. Applying this in the proof of Theorem 2.1a, given $r_n \|\hat{\beta} - \beta_0\|_\infty = O_p(1)$ and replacing everything with Frobenius norm for matrices with maximum row sum matrix norm, and l_2 norm for vectors with l_∞ norm, we have

$$r_n^* = O\left(\frac{r_n}{\|f_d(\beta_0)\|_\infty}\right). \quad (2.10)$$

Also part b) can be written in l_∞ norm as well.

5. What if we have $m = p?$ or $m > p$, and $m \rightarrow \infty$ as $n \rightarrow \infty$? Then all our results will go through as well, this is clear from our proof.

3 Examples

We now provide two examples that will highlight the contribution. First one is related to linear functions of estimators in part a), and the second one is related to risk of the large portfolios, and

part b).

Example 1.

Let us denote β_0 as the true value of vector $(p \times 1)$ of coefficients. The number of the true nonzero coefficients are denoted by s_0 , and $s_0 > 0$. A simple linear model is:

$$y_t = x_t' \beta_0 + u_t,$$

where $t = 1, \dots, n$, with u_t iid mean zero, finite variance error, and x_t is deterministic set of p regressors for ease of analysis.

The lasso estimator in a simple linear model is defined as

$$\hat{\beta} = \operatorname{argmin}_{\beta \in R^p} \sum_{t=1}^n \frac{(y_t - x_t' \beta)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^p |\beta_{j0}|,$$

where λ is a positive tuning parameter, and it is established that $\lambda = O(\sqrt{\frac{\log p}{n}})$. Corollary 6.14 or Lemma 6.10 of Buhlmann and van de Geer (2011) shows, for lasso estimators $\hat{\beta}$, with $p > n$

$$r_n \|\hat{\beta} - \beta_0\|_2 = O_p(1), \tag{3.1}$$

where

$$r_n = \sqrt{\frac{n}{\log p}} \frac{1}{\sqrt{s_0}}. \tag{3.2}$$

At this point, we will not go into detail such as what assumptions are needed to get (3.1), except to tell that minimal adaptive restrictive eigenvalue is positive, and noise reduction is achieved. Details can be seen in Chapter 6 in Buhlmann and van de Geer (2011).

The issue is what if the researchers are interested in the asymptotics of $D(\hat{\beta} - \beta_0)$, where $D : m \times p$ matrix. D matrix can be thought of putting restrictions on β_0 . We want to see whether $D(\hat{\beta} - \beta_0)$ has a different rate of convergence than $\hat{\beta} - \beta_0$. From our Theorem 2.1a, it is clear that $f_d(\beta_0) = D$. Assume that $\|D\|_2 \neq o(1)$. So

$$r_n^* \|D(\hat{\beta} - \beta_0)\| = O_p(1),$$

where

$$r_n^* = O\left(\frac{r_n}{\|D\|_2}\right). \quad (3.3)$$

We know from matrix norm definition:

$$\|D\|_2 = \sqrt{\sum_{i=1}^m \|d_i\|_2^2},$$

and $\|d_i\|_2$ is the Euclidean norm for vector d_i which is $p \times 1$, and

$$D = \begin{bmatrix} d'_1 \\ \vdots \\ d'_m \end{bmatrix}.$$

Basically in the case of inference, this matrix and vectors show how many of β_0 will be involved with restrictions. If we want to use s_0 elements in each row of D to test m restrictions, then $\|D\|_2 = O(\sqrt{s_0})$. Note that this corresponds to using s_0 elements in β_0 for testing m restrictions. So

$$r_n^* = \left(\sqrt{\frac{n}{\log p}}\right) \left(\frac{1}{s_0}\right), \quad (3.4)$$

which shows that even though we have fixed number of restrictions, m , using s_0 of coefficients in testing will slow down the rate of convergence, r_n^* by $\sqrt{s_0}$, compared with lasso estimators, rate of r_n . This can be seen by comparing (3.2) with (3.4).

Example 2.

One of the cornerstones of the portfolio optimization is estimation of risk. If we denote the portfolio allocation vector by w ($p \times 1$) vector, and the covariance matrix of asset returns by Σ , the risk is $\sqrt{(w'\Sigma w)}$. We want to analyze risk estimation error which is $(w'\hat{\Sigma}w)^{1/2} - (w'\Sigma w)^{1/2}$. $\hat{\Sigma}$ is the sample covariance matrix of asset returns. We could have analyzed risk estimation error with estimated weights, as in Fan et al (2015), $(\hat{w}\hat{\Sigma}\hat{w})^{1/2} - (\hat{w}\Sigma\hat{w})^{1/2}$, but this extends the analysis with more notation with the same results.

A crucial step in assessing the accuracy of risk estimator is given in p.367 of Fan et al (2015), which is the term $w'(\hat{\Sigma} - \Sigma)w$. . Just to simplify the analysis, we will assume iid, sub-Gaussian

asset returns. Also we will find the global minimum variance portfolio as in Example 3.1 of Fan et al (2015). So

$$w = \frac{\Sigma^{-1}1_p}{1_p' \Sigma^{-1}1_p},$$

where Σ is nonsingular, and 1_p is the p vector of ones. Assume $0 < \text{Eigmin}(\Sigma^{-1}) \leq \text{Eigmax}(\Sigma^{-1}) < \infty$, where $\text{Eigmin}(\cdot), \text{Eigmax}(\cdot)$ represents the minimal, maximal eigenvalues respectively of the matrix inside the parentheses. In Remark 3 of Theorem 3 in Caner et al (2016)

$$\|w\|_1 = O(\max_j \sqrt{s_j}), \quad (3.5)$$

where s_j is the number of nonzero cells in j th row of Σ^{-1} matrix, $j = 1, \dots, p$. Equation (3.5) represents case of growing exposure, which means we allow for extreme positions in our portfolio, since we allow $s_j \rightarrow \infty$, as $n \rightarrow \infty$. See that

$$\|\hat{\Sigma} - \Sigma\|_\infty = O_p(\sqrt{\frac{\log p}{n}}), \quad (3.6)$$

by van de Geer et al. (2014). Then clearly by (3.5)(3.6)

$$|w'(\hat{\Sigma} - \Sigma)w| \leq \|w\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty = O_p(\max_{1 \leq j \leq p} s_j \sqrt{\log p/n}). \quad (3.7)$$

This means taking $\beta_0 = w'\Sigma w, \hat{\beta} = w'\hat{\Sigma}w$, so $m = 1$ in Theorem 2.1b,

$$r_n |w'(\hat{\Sigma} - \Sigma)w| = O_p(1), \quad (3.8)$$

where

$$r_n = \sqrt{\frac{n}{\log p}} \frac{1}{\max_{1 \leq j \leq p} s_j}. \quad (3.9)$$

But the main issue is to get risk estimation error, not the quantity in (3.8). To go in that direction see that

$$O(\max_{1 \leq j \leq p} s_j) = \|w\|_1^2 \text{Eigmin}(\Sigma) \leq |w'\Sigma w| \leq \|w\|_1^2 \text{Eigmax}(\Sigma) = O(\max_{1 \leq j \leq p} s_j), \quad (3.10)$$

where we use (3.5) and $Eigmax(\Sigma) < \infty$, $Eigmin(\Sigma) > 0$.

Note that risk is $f(\beta_0) = (w'\Sigma w)^{1/2}$, and $f_d(\beta_0) = (w'\Sigma w)^{-1/2} = O((\max_j s_j)^{-1/2})$ in Theorem 2.1b. So $f_d(\beta_0) = o(1)$, since we allow $s_j \rightarrow \infty$. Then apply our delta theorem, Theorem 2.1b here

$$r_n[(w'\hat{\Sigma}w)^{1/2} - (w'\Sigma w)^{1/2}] = o_p(1), \quad (3.11)$$

Now we see that rate of convergence in risk estimation is faster in (3.11), compared to (3.8)-(3.9).

REFERENCES

- Abadir, K. and J.R. Magnus (2005). *Matrix Algebra*. Cambridge University Press. Cambridge.
- Buhlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data*. Springer Verlag, Berlin.
- Caner, M., E. Ulasan, L. Callot, and O. Onder (2016). "A relaxed approach to estimating large portfolios and gross exposure," arXiv: 1611.07347v1.
- Fan, J., Y. Liao, and X. Shi (2015). "Risks of large portfolios," *Journal of Econometrics*, 186, 367-387.
- Horn, R.A. and C. Johnson (2013). *Matrix Analysis*. Second Edition. Cambridge University Press, Cambridge.
- Van de Geer, S., P. Buhlmann, Y. Ritov, and R. Dezeure (2014). "On asymptotically optimal confidence regions and test for high-dimensional models," *The Annals of Statistics*, 42, 1166-1202.
- Van der Vaart, A.W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Appendix

Proof of Theorem 2.1a. The first part of the theorem shows why the classical proof of delta theorem will not work in high dimensions. However, this is not a negative result since it will guide us through the second part which shows us the solution.

Part 1. First by differentiability, and define $l(\cdot)$ as a vector function, via p.352 of Abadir and Magnus (2005) $l(\cdot) : D \subset R^p \rightarrow R^m$

$$\|f(\hat{\beta}) - f(\beta_0) - f_d(\beta_0)[\hat{\beta} - \beta_0]\|_2 = \|l(\hat{\beta} - \beta_0)\|_2. \quad (3.12)$$

and

$$\frac{\|l(\hat{\beta} - \beta_0)\|_2}{\|\hat{\beta} - \beta_0\|_2} = o_p(1), \quad (3.13)$$

where we use Lemma 2.12 of van der Vaart (2000). Then with $\|\hat{\beta} - \beta_0\|_2 = o_p(1)$, (3.13) implies

$$\|l(\hat{\beta} - \beta_0)\|_2 = o_p(1). \quad (3.14)$$

Since we are given $r_n\|\hat{\beta} - \beta_0\|_2 = O_p(1)$, by (3.13)(3.14)

$$r_n\|l(\hat{\beta} - \beta_0)\|_2 = o_p(1). \quad (3.15)$$

By (3.12)-(3.15)

$$\|r_n[f(\hat{\beta}) - f(\beta_0)] - r_n[f_d(\beta_0)][\hat{\beta} - \beta_0]\|_2 = o_p(1). \quad (3.16)$$

But this is the same result as in regular delta method. (3.16) is mainly a simple extension of Theorem 3.1 in van der Vaart (2000) to Euclidean spaces so far. However the main caveat comes from derivative matrix $f_d(\beta_0)$ which is of dimension $m \times p$. The rate of the matrix plays a role when $p \rightarrow \infty$ as $n \rightarrow \infty$. For example, both $r_n[f_d(\beta_0)][\hat{\beta} - \beta_0]$ and $r_n[f(\hat{\beta}) - f(\beta_0)]$ may be diverging, but $r_n\|\hat{\beta} - \beta_0\| = O_p(1)$. Hence the delta method is not that useful if our interest centers on getting rates for estimators as well as functions of estimators that converge. In the fixed p case, this is not an issue, since the matrix derivative will not affect the rate of convergence at all, as long as this is bounded away from zero, and bounded from above. Note that boundedness assumptions may not

be intact when we have $p \rightarrow \infty$, as $n \rightarrow \infty$. Next part shows how to correct this problem.

Part 2. From differentiability, using p.352 of Abadir and Magnus (2005), or proof of Theorem 3.1 in van der Vaart (2000)

$$f(\hat{\beta}) - f(\beta_0) = f_d(\beta_0)[\hat{\beta} - \beta_0] + l(\hat{\beta} - \beta_0).$$

Putting the above in Euclidean norm, and using triangle inequality

$$\begin{aligned} \|f(\hat{\beta}) - f(\beta_0)\|_2 &= \|f_d(\beta_0)[\hat{\beta} - \beta_0] + l(\hat{\beta} - \beta_0)\|_2 \\ &\leq \|f_d(\beta_0)[\hat{\beta} - \beta_0]\|_2 + \|l(\hat{\beta} - \beta_0)\|_2 \end{aligned}$$

Next, multiply each side by r_n , and use (3.15)

$$r_n \|f(\hat{\beta}) - f(\beta_0)\|_2 \leq r_n \|f_d(\beta_0)[\hat{\beta} - \beta_0]\|_2 + o_p(1). \quad (3.17)$$

Then apply matrix norm inequality in (2.1) to the first term on the right side of (3.17)

$$r_n \|f_d(\beta_0)[\hat{\beta} - \beta_0]\|_2 \leq r_n [\|f_d(\beta_0)\|_2] \left[\|\hat{\beta} - \beta_0\|_2 \right]. \quad (3.18)$$

Substitute (3.18) into (3.17) to have

$$r_n \|f(\hat{\beta}) - f(\beta_0)\|_2 \leq r_n [\|f_d(\beta_0)\|_2] \left[\|\hat{\beta} - \beta_0\|_2 \right] + o_p(1). \quad (3.19)$$

Now divide each side by $\|f_d(\beta_0)\|_2$, since $\|f_d(\beta_0)\|_2 > 0$, and $\|f_d(\beta_0)\|_2 \neq o(1)$,

$$\frac{r_n}{\|f_d(\beta_0)\|_2} \|f(\hat{\beta}) - f(\beta_0)\|_2 \leq r_n \left[\|\hat{\beta} - \beta_0\|_2 \right] + o_p(1). \quad (3.20)$$

By Assumption since $r_n \left[\|\hat{\beta} - \beta_0\|_2 \right] = O_p(1)$, we have

$$\frac{r_n}{\|f_d(\beta_0)\|_2} \|f(\hat{\beta}) - f(\beta_0)\|_2 = O_p(1) + o_p(1). \quad (3.21)$$

So the result is derived, by noting that the new rate of convergence for the function of estimators is

$$r_n^* = O\left(\frac{r_n}{\|f_d(\beta_0)\|_2}\right).$$

Q.E.D.

Proof of Theorem 2.1b. Here we use (3.19), and since $\|f_d(\beta_0)\|_2 = o(1)$, and

$$r_n\|\hat{\beta} - \beta_0\|_2 = O_p(1),$$

we have

$$r_n\|f(\hat{\beta}) - f(\beta_0)\|_2 \leq o_p(1) + o_p(1) = o_p(1).$$

Q.E.D.